ANNUAL PROGRESS REPORT

COMPUTER CLASSIFICATION OF DOCUMENTS

A paper presented to the FID/IFIP Conference
in Rome, Italy, on June 15, 1967

J. H. Williams, Jr.

CONTRACT NONR 4456(00)

Submitted to

Information Systems Branch
Office of Naval Research
Department of the Navy
Washington, D. C. 20360

Federal Systems Division
International Business Machines Corporation
Gaithersburg, Maryland 20760

25

COMPUTER CLASSIFICATION OF DOCUMENTS*

J. H. Williams, Jr.
Federal Systems Division
International Business Machines Corporation
Gaithersburg, Maryland 20760

Published in the
Proceedings of the FID/IFIP Conference 1967
on Mechanized Information Storage Retrieval and
Dissemination, Rome, Italy, June 1967

*Joint support for this effort has been provided by RADC, AFSC under Contract AF 30(602)-4161, RADC-TR 67-191; ONR under Contract NONR 4456(00) and IBM's IRAD program and its International Patent Operations.

COMPUTER CLASSIFICATION OF DOCUMENTS

Abstract

Classification of documents involves three distinct major processes. The first two processes of defining a structure of categories and determining a basis for the classification decision are usually performed by a classificationist, while the third process of classifying documents into categories is performed by a classifier. The objectives of our approach is to develop computer techniques to perform the second and third processes.

Previous experiments indicate that all terms do not need to be retained for the classification process, and computationally it would be impractical to do so. Therefore, a word selection measure is employed to delete those terms that rarely occur and those that have a low conditional probability of occurring in a category. A set of sample documents known to belong to each category is used to estimate the mean frequency, the within category variance and the between category variance of the remaining terms. These statistics are then employed to compute discriminant functions which provide weighting coefficients for each term.

A new document is classified by counting the frequencies of the selected terms occurring in it, and weighting the difference between this vector of observed frequencies and the mean vector of every category. The probability of membership in each category is computed and the document is assigned to the category having the highest probability. For applications in which assignment to one category is not desirable, the probabilities can be used to indicate multi-category assignment.

A thesaurus capability allows the following types of words to be considered equivalent: inflected words, compound words, and semantically similar words with different orthographic spellings. Since the technique is based on statistical measures, it can classify documents written in any language provided a sample set of documents in that language is available.

Experiments have been conducted on several English data bases, and a further experiment is being conducted on a German data base. Classification results in a recent experiment have ranged from 73 to 95 percent.

INTRODUCTION

Both indexing and classification accomplish the same process of assigning a tag to a document, and have the same objective of retrieving relevant documents on the basis of their tags. A classification system, in addition to providing tags, also provides an organization of the tags based on the classification structure. For some applications assignment to a category does not provide a sufficiently fine partition of a collection for effective retrieval. Therefore, we have developed a two-stage technique consisting of searching for relevant categories and then querying within those categories for relevant documents.

Classification of documents involves three distinct major processes. The first two processes of defining a structure of categories and determining a basis for the classification decision are usually performed by a classificationist, while the third process of classifying documents into categories is performed by a classifier. The objective of our approach is to develop computer techniques to perform the second and third processes. Because a particular subject field may be partitioned in many ways depending upon the point of view and needs of the user, we believe that the classificationist's first process must be influenced by the needs of his organization. Therefore, rather than attempt to define categories or cluster documents statistically to determine a mathematically optimum partition, we accept the user's structure and start our technique with a sample of documents known to belong to each category. Each category in the structure is considered to be a node in a tree, and all nodes below that node are its subcategories.

Our current computer programs perform the second and third processes. The first set of programs attempts to detect a pattern among the documents and then select and weight a subset of words to form a basis for the classification decision. These classificationist programs are used only when the system is initiated or revised, whereas the second set of programs are used periodically to classify new documents.

The classifier programs could be modified to not only classify new documents but also store frequency counts on all words observed in the documents, along with the categories to which it was assigned. Periodically (or on demand), comparisons could then be made between statistics collected from the new documents and the statistics collected on the original documents. When a significant difference occurred in any one of the statistics, an output could be generated for perusal by the classificationist.

Information required for the addition of categories can be obtained readily be observing an increase in the arrival rate of new items. Information for the deletion of categories can be obtained by observing either a decrease in the arrival rate of documents in a specific category or a decrease in the arrival rate of terms in the discriminating subset. In our technique, the categories are actually defined by only a small subset of terms. By changing the terms within the subset, the definition of the categories will be changed. Statistics indicating the potential discriminating power and the coverage of each term will be maintained separately for each category. Thus the need for creating an inter-disciplinary category can be observed when the arrival rate of a term increases simultaneously in several apparently unrelated categories.

CLASSIFICATION PROCEDURE

A user selects a classification structure and a sample of documents known to belong to each category. The text of these documents (or abstracts) is entered into the computer. A word frequency program counts the frequency of each word type for each category.

Previous experiments indicate that all word types do not need to be retained for the classification process, and computationally it would be impractical to do so. Ideally, words selected to represent the categories should occur in one and only one category. However, there are usually only a few words in any data base that occur in one and only one category, and these words do not necessarily occur in every document. Therefore, a word selection statistic is needed to identify words approximating this condition, and to select a subset of words to form the basis of the classification decision. The statistic chosen is the log of the ratio of the relative frequency of a word in a category to the relative number of documents in that category, and this is computed for each word in the category.

For each word, the value of this statistic is compared across all categories, and a particular word is placed in the list of that category in which it has the most positive value. After all words have been placed in a category, the word list for each category is arranged in descending values of the statistics.

Finally, to represent each category in the structure, words are selected according to two criteria. Words must not only have a high word selection statistic value, but they must also occur in some specified minimum number of documents in the category. Thus, the latter criterion is needed to ensure that the subset of words selected will provide a significantly high percentage coverage. The words that satisfy both requirements will be called discriminating words. They form the basis for all classification decisions to be made at the branch point for which they were developed.

At present our computer programs will accept only 100 of these discriminating words. In order to obtain maximum coverage from this relatively small set, the following thesaurus techniques have been incorporated:

(1) Various inflections of a word are combined with its root word

(2) Compound words are combined with their root words

(3) Synonyms and related words are tagged with the same internal word number.

These techniques effected an increase in coverage of approximately 200 more words.

The sample set of digests for each category are again processed to compute the mean frequency of each discriminating word for each category, the pooled within-category dispersion, W, and the among-category dispersion, A. The optimum set of weighting coefficients is found by solving the determinantal equation, $|W^{-1} \ A - 1\lambda| = 0$, for its eigenvalues, λ. The eigenvalues are then used to compute eigenvectors whose elements are the desired weighting coefficients. The number of non-zero eigenvalues of the determinantal equation is at most equal to the smaller of the number of categories minus one or the number of variables. Thus, our technique is independent of the number of categories. If a group contained ten categories, nine eigenvalues would be found which would provide nine sets of weighting coefficients for each word.

The eigenvalue solution also provides the basis of an orthogonal discriminant (classification) space. The eigenvectors are used to transform each category mean and dispersion from the original 100-variable space to a reduced classification space.

A new document is classified by counting the frequencies of the discriminating words occurring in it, transforming this frequency vector to the classification space (weighting its words) and comparing it with the mean vector of every category. The probability of membership in each category is computed and the document is assigned to the category having the highest probability. For applications in which assignment to only one category is not desirable, the probabilities for each category may be stored for future retrieval.

WORD FREQUENCY PROGRAM

In conjunction with this project a generalized character and word frequency program has been developed for the System/360 computer. These programs (1) are being used in the computer classification experiments and can be used independently for any language analysis study involving the statistical and morphological behavior of character strings or items in narrative text.

The S/360 program is written in FORTRAN IV and can be easily adapted to a 360 model available to the user. The program provides numerous user options concerning the definition of a countable item (e.g., a single character or a character string, which may or may not be a word; a "word" may be specified as any string of characters between delimiters such as comma, space, period, or any combination thereof), the definition of the textual units over which frequencies are to be subtotaled (e.g., sentence, paragraph, and/or document), the types of data to be output, and the machine configuration to be used.

The modular program design provides subroutines that perform functions basic to all applications and subroutines that perform optional functions specified by the user. It also allows for the incorporation of new programs to be written by the user to perform additional optional functions. The basic subroutines incorporated in the program perform the input and item identification, dictionary building, merging, and frequency output functions. The program-provided optional subroutines perform the concordance, special item check, summary output, growth rate, and detailed frequency print functions. Some user-provided optional programs could perform pre-processing, interval definition, encoding, word use tagging, and special action on specific word functions.

Detailed output available for each item includes the item itself, its character length, its frequency in absolute and percentage form, the location of its first occurrence and the number of textual units in which it appeared. Summary outputs available are vocabulary growth rate, distribution, item types by initial character, item types by string length, item tokens by string length, and a concordance of items, tags, interval identification and sequential position within interval. Each of these outputs may be obtained for any or all textual units.

CLASSIFICATION EXPERIMENTS

A series of experiments have been conducted to demonstrate the generality of the technique on data bases from various disciplines and on data bases in the English and German languages.

The experiments have also provided information on ranges of values of significant parameters, which are necessary to determine the effectiveness of the technique on a particular data base.

Table 1 contains a summary of the results and conditions of four experiments. The earliest work (2) consisted of a computer evaluation of the form of the classification equations proposed by Edmondson and Wyllys (3) and classification experiments on computer abstracts of the same type used by Maron (4) and Borko (5). These experiments (6) indicated that better results could be achieved by using a subset of all the words occurring in a document collection and by weighting words according to their discrimination ability rather than treating each word equally in the classification decision.

Many statistical techniques exist for the classification of a random observation into one of two populations. However, not until recently have techniques been developed for classifying observations into many categories. A survey of the techniques has indicated that multiple discriminant functions appear to be the best statistical technique for document classification. The functions not only provide weighting coefficients that reflect a word's discriminating ability but they also offer the optimum classification decision rule (7) when the multivariate data is normally distributed. Data from the solid state

Table 1. Summary of Classification Experiments

Subject Field	Computer	Legal	Solid State	Computer
Language	English	English	English	German
Type of document	Abstract	Document	Abstract	Abstract
% agreement of computer with original classification				
Sample documents	--	98%	88%	94%
Test documents	67%	74%	79%	90%
Source of original classifications	CCC*	West	CCC*	GFRPO**
# Documents available	400	5000	2754	5000
# Documents included in experimental structure	400	885	1743	2097
# Sample documents in each category	15, 75	20-48	35, 70, 140	141-937
# Levels in experimental structure	2	2	2	3
# Groups in experimental structure	5	2	2	3
# Categories in a group	4, 5	4, 5	3	2, 3
Total # of categories	24	9	6	7
# Discriminating words	20	48	48	100
Average length of document	90	1000	90	30
Average # of discriminating words in document	--	10	6	3
Thesaurus capability	No	No	Yes	Yes

*CCC is Cambridge Communication Corporation.

**GFRPO is German Federal Republic Patent Office experiment plotted in Figure 1 indicates that the coordinates of documents in the classification space appear to be bivariate normally distributed since they are enclosed by an ellipse. The data in the upper plot is based on the sample documents used to generate the system whereas the lower plot consists of new documents that are presented to the system for classification. An ellipse indicating the 99% contour line should enclose the observations of a sample or population with a 99% probability. Since the plot of sample documents is similar to the plot of an independent set it has been concluded that the distribution of the sample is an adequate representation of the population distribution, and they are both normally distributed.

Multiple discriminant functions have been used in each of the succeeding experiments. The legal experiment demonstrated that documents longer than abstracts could be handled. The documents ranged in length from 500 to 5000 words. The longer documents performed better than the shorter ones. The legal profession requires two different types of searches on the same data base. They may wish to find a document relevant to points of law in the case at hand or they may wish to find a document relevant to the facts in the case at hand. Thus, the same data base must be partitioned and classified from two points of view. This was

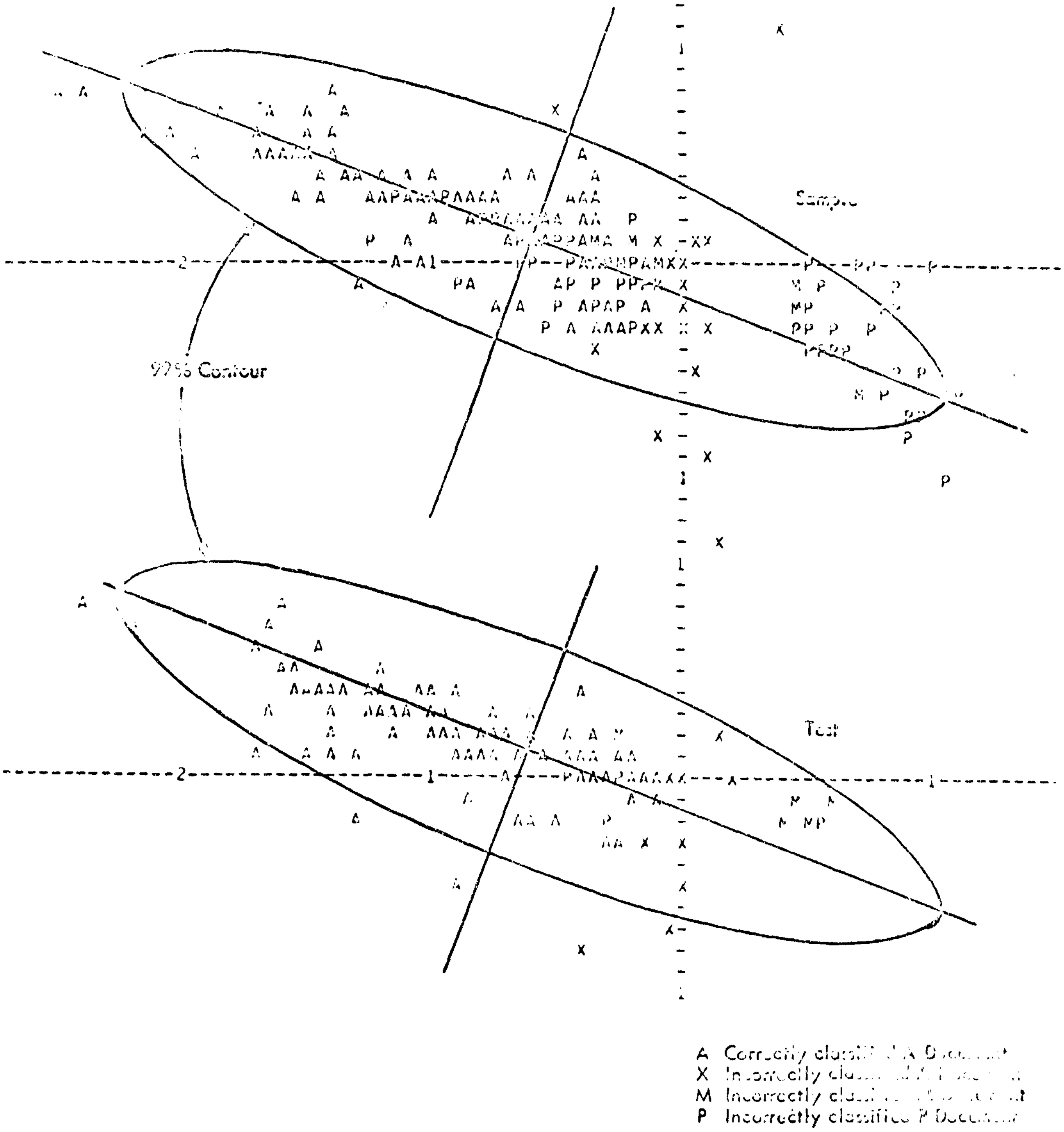

Figure 1. Solid-State Document Classification Space

accomplished by first selecting a subset of law words and a subset of fact words, and secondly, classifying each document twice. The two resulting files are independent and searches may be addressed to either or both files. No significant difference was observed in classification performance between the two classification systems.

The solid state experiment (8) provided information on the significant parameters affecting classification performance. The parameters studied were the number of sample documents required to define a category, the length of documents, the interrelationships of the number of sample documents and their lengths, the relation of the number of word types in a document to the number of categories assigned to it, levels in a structure, homogeneity of categories, and the number of discriminating types occurring in a document.

The number of sample documents required to form the basis of the classification decision appeared to be an important parameter. Experiments were conducted with 35, 70, and 140 sample documents per category. As the number of sample documents increased the performance on the sample decreased whereas the performance on the independent test set increased. When performance on both sets converge, the maximum performance of the system can be determined (if no other parameters are changed) and it can be concluded that the sample is representative of the population.

Perfo~~r~~mance is not wholly dependent on the number of sample documents but rather on the total words in the sample. Thus, fewer longer documents may be required to reach a stable point as in the legal experiment where as few as twenty sample documents were used having a length of 500 to 5000 words. The difference between the solid state sample and the test results is much less than the difference in the legal results as shown in Table 1.

The classification procedure in a structure consisting of many levels and many subcategories involves an independent decision at each branch point (node) in the structure. For a structure containing five levels, five classification decisions are made. The basis for a decision at one level is independent of the basis at another level. The basis at each node is determined by the sample documents within that node and the discriminating subset of words derived from those documents. A different discriminating subset is used at each node. Words may or may not be members of subsets at various nodes, depending upon their discriminative ability at a node. A solid state experiment indicated that there was no degradation in performance at a lower level when the number of sample documents was held constant.

The latest experiment was performed on a set of patent abstracts concerning computer circuits supplied by the IBM Germany Patent

Department. The abstracts written in the German language, were originally classified by the German Federal Republic Patent Office. Samples of documents were randomly selected from each category to derive the discriminating word subsets and to form the basis of the classification decision. To preserve the a priori distribution of documents over the categories, two-thirds of the documents available in each category were selected for the sample set. This yielded a range from 141 to 937 documents per category, the categories at the lowest levels having the fewest documents.

Language translation programs were unnecessary for the technique to operate on the German language data base. The programs compute statistics on the words contained in the sample documents.

A thesaurus capability incorporated with the solid state experiment was expanded for the German experiment. As the discriminating words are being selected, inflected forms of a word are considered equivalent to its root word, compound words occurring with similar discriminative power in the same category are considered equivalent (E'NGANG, EINGANGSKLEMME, EINGANGSSIGNAL, EINGANGSIMPULS), and words having the same discriminative power in the same category occurring with different orthographic representations are considered equivalent (FLIP-FLOP, MULTI-VIBRATOR).

Since a different discriminating word set exists for each group, the thesaurus relationships hold only for that group. This provides a solution for the arduous and paradoxical task of constructing a single thesaurus for a given data base. It allows contextual relationships dependent on the particular subject group. If the word "pitch" occurs in three different groups it can be related to different words in each group: throw (sports), level (music), tip (dynamics).

The technique was tested at the second, third, and fifth level of detail in the German patent structure. The fifth level consisted of deciding within the pulse circuitry group whether the circuit generated pulses, switched pulses or counted pulses. The overall performance yielded 90% agreement with the original categories for the independent test set and 94% for the sample set.

Successful computer classification experiments have been performed on four data bases involving over 5000 documents in two languages. The experiments have yielded considerable data on the significant classification parameters which can be used to design computer classification systems and improve their performance. Consideration has been given to problems of changing technology and the need for updating classification structures, reclassifying documents and recognizing the arrival of new terms.

A two-stage searching technique consisting of searching for relevant categories and searching for relevant documents within a category based on a full text strategy is now under development. Documents are classified within a structure and a concordance of terms occurring in each document is prepared. A query is presented to the system in the form of a statement of the problem written in natural text approximately a paragraph long. The query is classified into one or more categories. Then a fine search is made with a term by term comparison of the query and each document in the category.

ACKNOWLEDGMENTS

The author wishes to extend his appreciation for the valuable help provided by his colleagues and in particular to Mr. Matthew P. Perriens for his continuous significant contributions, to Mrs. Kay Mandzak Baker for her astute research which led to the application of multiple discriminant functions to the document problem, and to Mr. F. T. Baker for the development of the generalized word frequency program.

REFERENCES

1. Baker, F. T., Johnson, G. L., Jones, M., Williams, J. H., Research on Automatic Classification, Indexing, and Extracting, NONR 4456(00), April 1966, AD 485188.

2. Meadown, Harriet R., Statistical Analysis and Classification of Documents, IRAD Task No. 0353, FSD, IBM, Rockville, Maryland, 1962.

3. Edmundson, H. P. and Wyllys, R. E., "Automatic Abstracting and Indexing--Survey and Recommendations," Communications of Association for Computer Machinery, Vol. 4, (1961), No. 5.

4. Maron, M. E., Automatic Indexing: and Experimental Inquiry, J. Assoc. Comp. Mach. 8, No. 3, 404-417 (1961).

5. Borko, H., and Mr. Bernick, Automatic Document Classification, J. Assoc. Comp. Mach. 10, No. 2, 151-162 (1963).

6. Williams, J. H., Results of Classifying Documents with Multiple Discriminant Functions, National Bureau of Standards' Symposium on Statistical Association Methods for Mechanized Documentation, Washington, D. C., April 1964.

7. Rao, C. Radhakrishna, Advanced Statistical Methods in Biometric Research, New York, Wiley & Sons, 1952.

8. Williams, J. H., Discriminant Analysis for Content Classification, AF 30(602)-35[illegible], RADC-TR-66-6, Griffiss AFB, New York, December 1965, AD 630-1[illegible]7

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Federal Systems Division
International Business Machines Corporation
Gaithersburg, Maryland 20760

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

RESEARCH ON AUTOMATIC CLASSIFICATION, INDEXING AND EXTRACTING

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Annual Progress Report

5. AUTHOR(S) *(Last name, first name, initial)*

Williams, Jr., John H.

6. REPORT DATE

December 1967

7a. TOTAL NO. OF PAGES

25

7b. NO. OF REFS

8

8a. CONTRACT OR GRANT NO.

NONR 4456(00)

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC. Other qualified users shall request copies of this report from the originator.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Information Systems Branch
Office of Naval Research
Dept. of the Navy, Washington, D. C.

13. ABSTRACT

Classification of documents involves three distinct major processes. The first two processes of defining a structure of categories and determining a basis for the classification decision are usually performed by a classificationist, while the third process of classifying documents into categories is performed by a classifier. The objectives of our approach is to develop computer techniques to perform the second and third processes.

Previous experiments indicate that all terms do not need to be retained for the classification process, and computationally it would be impractical to do so. Therefore, a word selection measure is employed to delete those terms that rarely occur and those that have a low conditional probability of occurring in a category. A set of sample documents known to belong to each category is used to estimate the mean frequency, the within category variance and the between category variance of the remaining terms. These statistics are then employed to compute discriminant functions which provide weighting coefficients for each term.

A new document is classified by counting the frequencies of the selected terms occurring in it, and weighting the difference between this vector of observed frequencies and the mean vector of every category. The probability of membership in each category is computed and the document is assigned to the category having the highest probability. For applications in which assignment to

ABSTRACT

Continuation Sheet
to Form DD 1473

one category is not desirable, the probabilities can be used to indicate multi-category assignment.

A thesaurus capability allows the following types of words to be considered equivalent: inflected words, compound words, and semantically similar words with different orthographic spellings. Since the technique is based on statistical measures, it can classify documents written in any language provided a sample set of documents in that language is available.

Experiments have been conducted on several English data bases, and a further experiment is being conducted on a German data base. Classification results in a recent experiment have ranged from 73 to 95 percent.

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Retrieval						
Subject Indexing						
Statistical Analysis						
Information Systems						
Documentation						
Libraries						
Indexes						
Decision Making						
Classification						
Word Association						
Correlation Techniques						
Dictionaries						
Vocabulary						
Pattern Recognition						
Information Sciences						
Automatic						
Indexing Terms						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through ________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through ________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through ________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

UNCLASSIFIED

Security Classification